A STUDY ABOUT HATE SPEECH RECOGNITION ACROSS MULTIPLE SOCIAL MEDIA

Hate-speech recognition

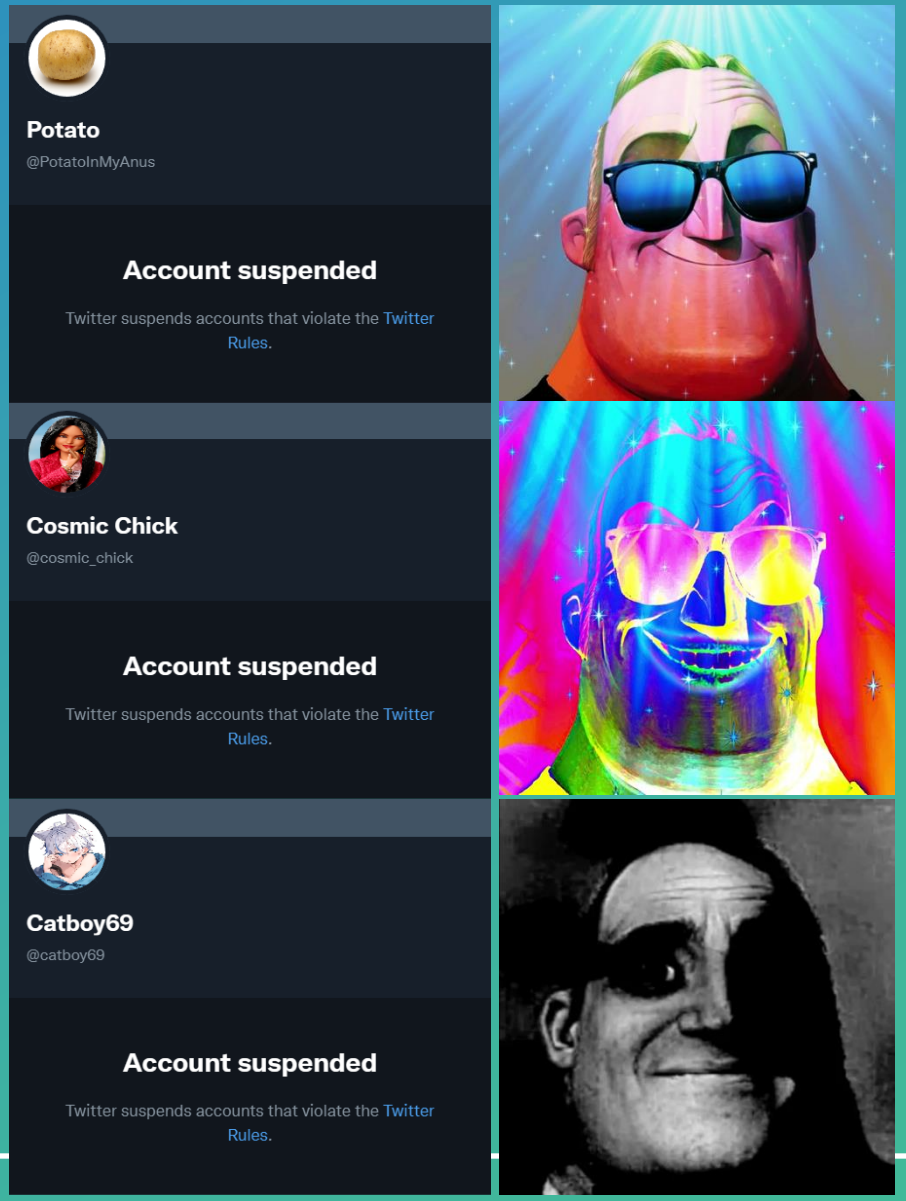

Catboy69
@catboy69
I love pineapple pizza!
1:32 PM · Jun 1, 2024
2 Retweets
13 Quote Tweets
15 Likes
Cosmic Chick @cosmic_chick · Jun 1
Everybody keeps on hating, but that shit ain't bad.
Potato @PotatoInMyAnus · Jun 1
Typical n***a, go hang yourself.

You

We probably want to stop here ☺

Unless you are Italian

Pov: Twitter launches AI hate-speech detector



On
note: h

writing showing prejudice
sexual orientation, etc.

intent, language.

sets

1

0

What (if any) is the difference between textual hate-speech detection across various social media?

Goals:

- Develop deeper understanding of hate speech and related challenges
- Discover possibilities of combining different content origins (social media) to enrich detection
- Develop generalizable and accurate hate-speech detection pipelines

Current Literature VS Unexplored Areas

Various ML techniques to deal with identifying hate-speech:

- Supervised Learning (SVM, DT, NN, LSTM)
- Unsupervised Learning (Clustering, AD)
- Transformer-based: (ro)BERT(a), GPT-x, LLaMA

Generally Transformer-based > others

Various studies into detecting hate-speech across particular social media platforms.

- Scarce up-to-date literature due to very rapidly evolving field
- Scarce literature comparing hate-speech detection performance across different social media platforms
- Scarce literature comparing platform-specific hate speech model performance with cross-platform performance.

Methodology

Data collection:

- Kaggle datasets

Data preprocessing:

- Standardizing and cleaning data across datasets
- Data balancing

Hate speech detection:

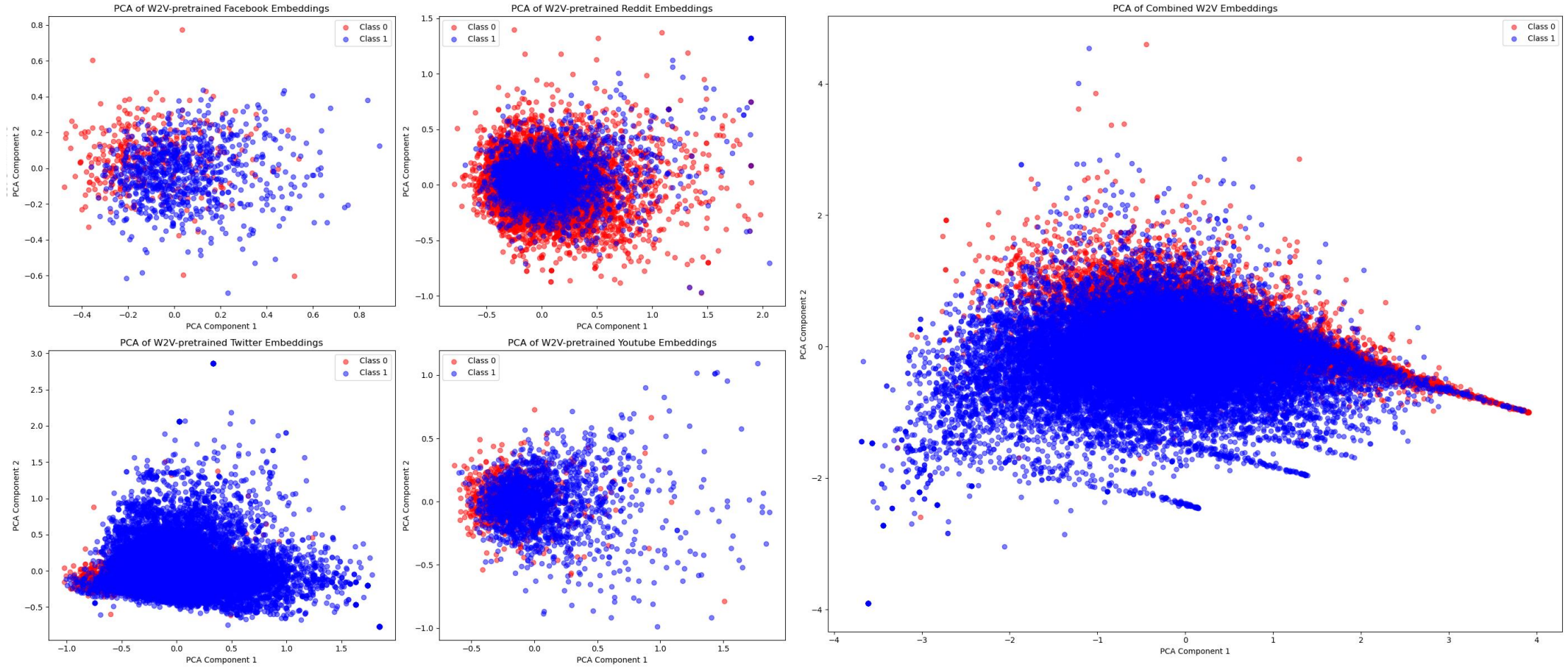
- Random Forest based on TF-IDF or W2V
- GPT-3.5 and BERT

Data Preparation and Cleaning

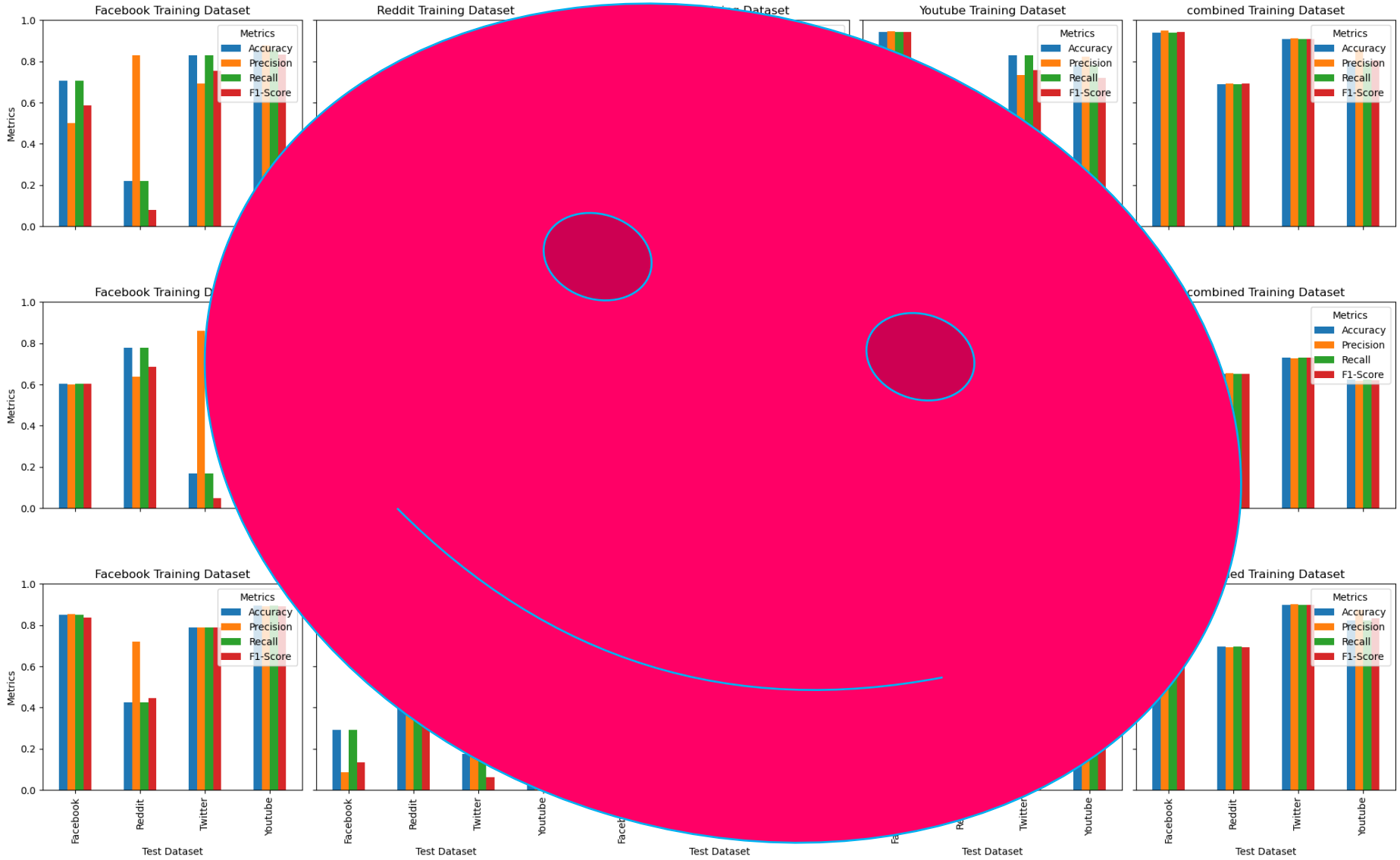
- Standardized formats of the csvs
- Split per platform
- Remove special annotations (i.e. RT for retweets
- Removed empty, “[deleted]” ,”[removed]”, blanks, nulls, zeroes
- Removed handles, usernames, emails, links, very short messages
- Removed duplicates with different labels

- Balanced datasets with undersampling / oversampling

Basic ML and embeddings



Experiment: tf-idf

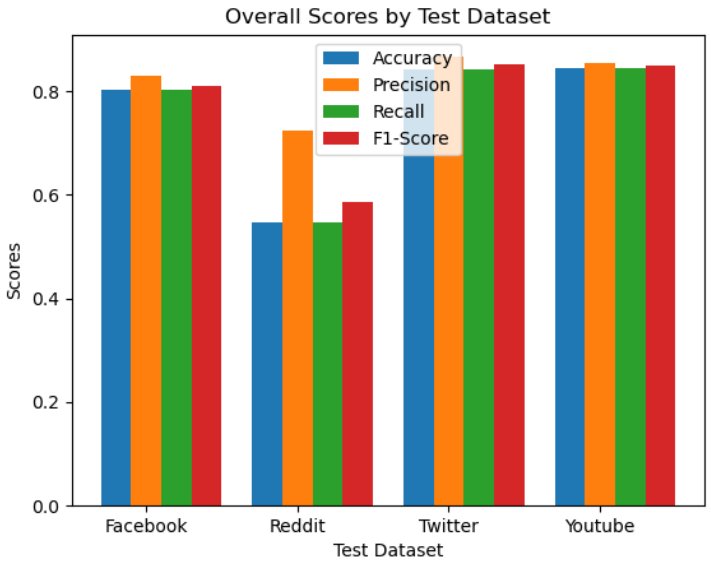


Basic Lessons learned

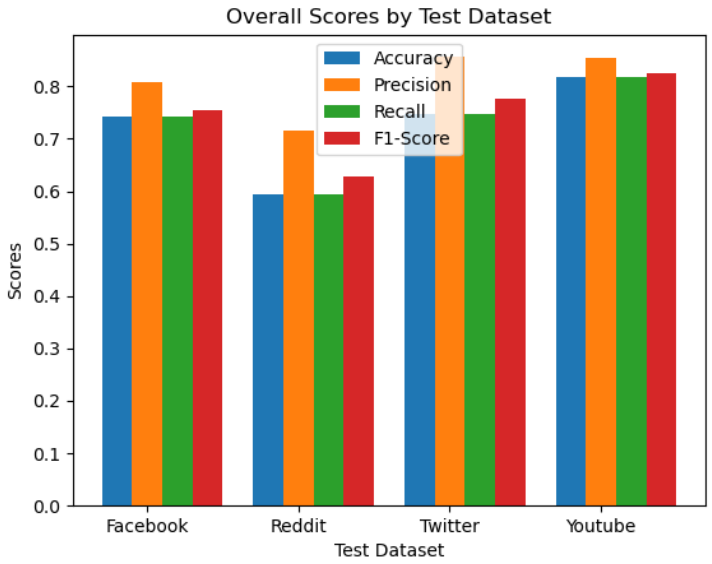
- Platform-specificity: best results for the same medium
- Combined datasets are op
- Nothing beats data quantity and diversity
- Low dataset quality in the field
- Weak embeddings, need for complex pretrained models

GPT for hate-speech detection

Simple prompt

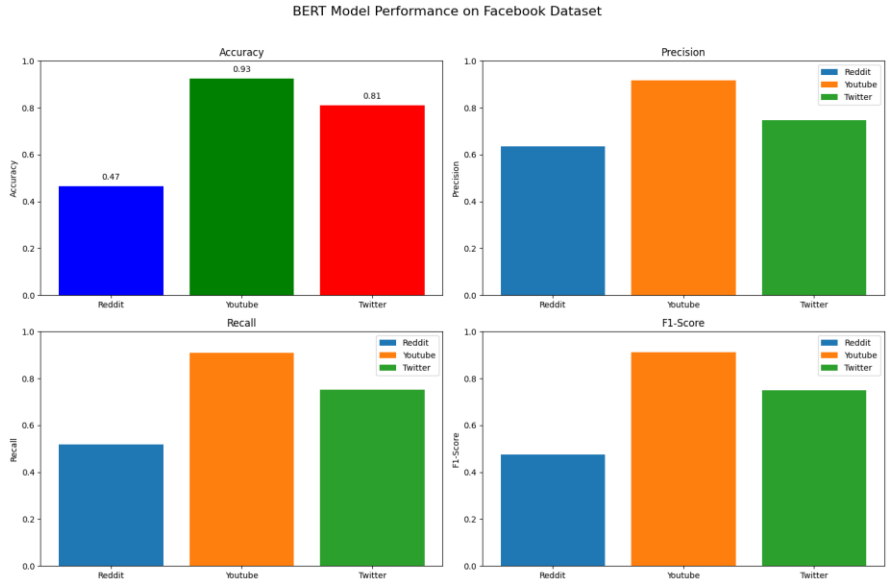


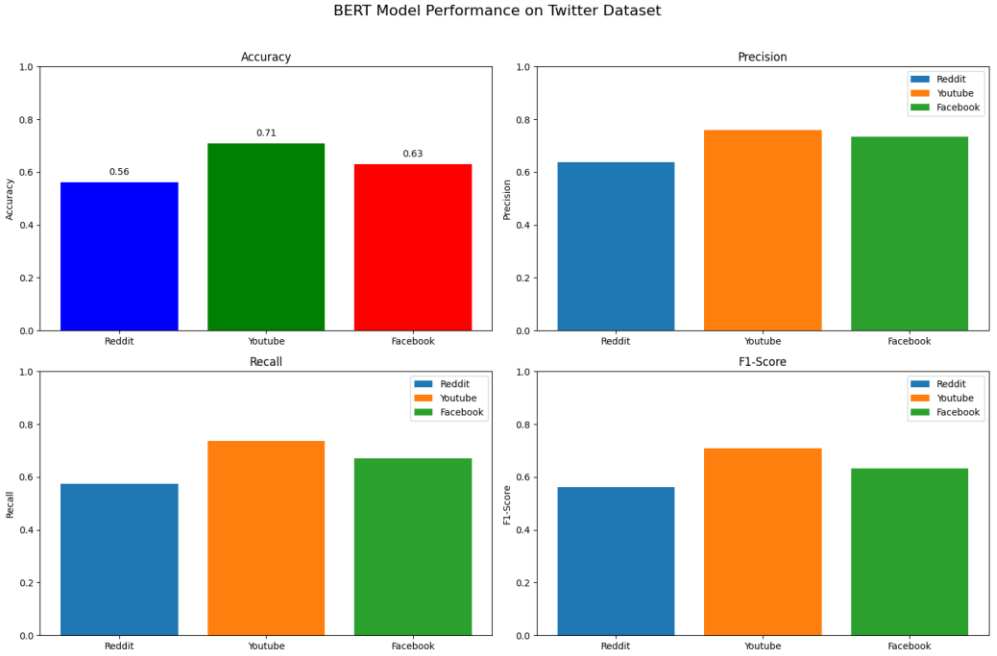
Prompt engineering

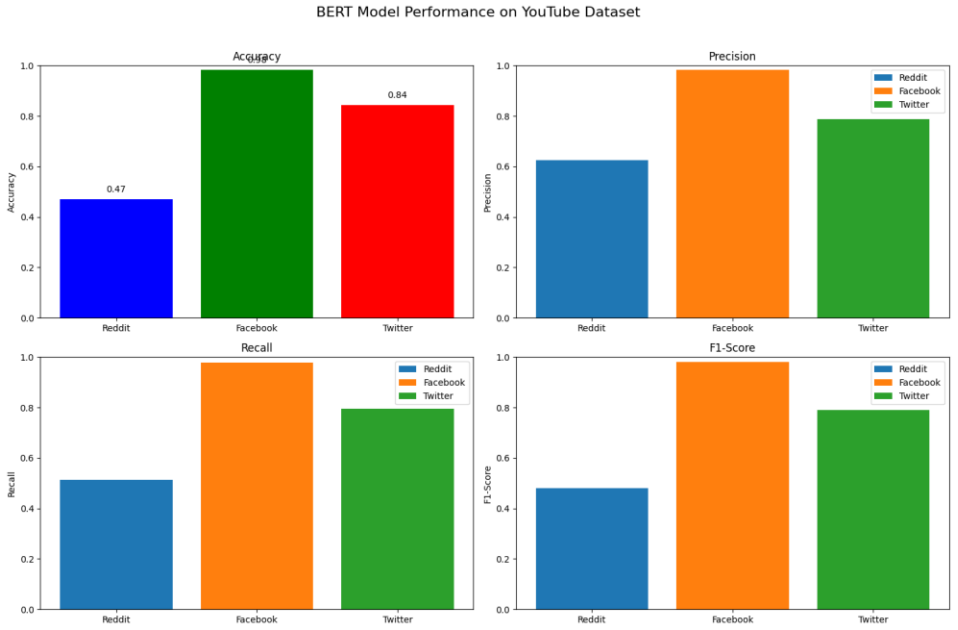


- Zero-shot sucks
- One-shot/Few-shots is impractical
- Fine-tuning is expensive/impractical
- GPT-4o is even more expensive (these 2 tests would net us 70 bucks)
- Overall, local models are better

BERT results







Preliminary results:
Significant drop in accuracy for reddit suggest, bert model struggles with the specific characteristics of reddit text

Looking at the comparisons training the model on twitter preforms marginally better on reddit than training on other platforms data.

We plan to still compare with Roberta.

Discussion / Conclusion

- The better results with the pre-trained RF W2V-based model in comparison to both the raw and (attempt at a) fine-tuned GPT-3.5 models, suggesting that new != better. (GPT-4 might be better, but very expensive)
- The combined model had on average the best overall results. This suggests that hate-speech models benefit from variety across platforms.
- Reddit had by far the worst results in hate-speech detection. That might be due to the usually larger comments and more niched language used.

- Using pre-trained 'traditional' ML techniques can have better results than new out-of-the-box transformer-based models without advanced fine-tuning.
- This study supports the general idea that more data (almost) always beats domain-specific text.
- Reddit seems to be the platform with hate-speech that is the hardest to detect.

Question Time